BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Network-enabled gene expression analysis

David Edwards*, Lei Wang and Peter Sørensen

Abstract

Background: Although genome-scale expression experiments are performed routinely in biomedical research, methods of analysis remain simplistic and their interpretation challenging. The conventional approach is to compare the expression of each gene, one at a time, between treatment groups. This implicitly treats the gene expression levels as independent, but they are in fact highly interdependent, and exploiting this enables substantial power gains to be realized.

Results: We assume that information on the dependence structure between the expression levels of a set of genes is available in the form of a Bayesian network (directed acyclic graph), derived from external resources. We show how to analyze gene expression data conditional on this network. Genes whose expression is directly affected by treatment may be identified using tests for the independence of each gene and treatment, conditional on the parents of the gene in the network. We apply this approach to two datasets: one from a hepatotoxicity study in rats using a PPAR pathway, and the other from a study of the effects of smoking on the epithelial transcriptome, using a global transcription factor network.

Conclusions: The proposed method is straightforward, simple to implement, gives rise to substantial power gains, and may assist in relating the experimental results to the underlying biology.

Background

Although genome-scale expression experiments are performed routinely in biomedical research, understanding the data they generate remains a major challenge. A widely used approach to relate such data to biology is *gene set enrichment analysis* [1]. This shifts focus from the regulation of individual genes to the regulation of gene sets, i.e. pre-defined sets of genes that share a common biological function or feature such as chromosomal location. One motivation for the approach is that if genes in a set are closely related, they can be expected to have similar expression patterns, and exploiting this may increase power. Another is that analysis at the level of individual genes may often be difficult to interpret biologically, whereas higher-level analyses hopefully allow a more coherent picture of the involved biological processes to emerge. Many methods of gene set enrichment analysis have been proposed (see [2] for a useful survey) and there exists a burgeoning literature on the topic. Very commonly, gene sets are defined on the basis on gene ontology [3] or pathway membership [4], using resources such as the Kyoto Encyclopedia of Genes and Genomes [5].

Recently there has been intense interest in methods that build on the information in biological networks, that is to say, methods that exploit the topology rather than just the set of genes in the network. We briefly summarize some of the methods proposed.

One approach extends gene set enrichment analysis by defining scores that build on network topology. For example, gene set scores that can be expressed as sums of pairwise weights between genes in the set may be modified by weighting gene pairs by the inverse of their path distance in the network [6]. Significance tests using the scores are then constructed, to assess whether the pathway is regulated. Similarly, scores that take the proximity of differentially expressed genes in the pathway into account have been proposed [7,8].

Another approach makes explicit use of network models for the expression data. In [9], fold changes are discretized into two levels: differentially expressed or equally expressed, and a Markov random field model (an undirected graphical model) for the fold changes whose parametrization utilizes pathway topology is adopted. Bayesian inference is used in which prior distributions for

*Correspondence: David.Edwards@agrsci.dk
Department of Molecular Biology and Genetics, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

the parameters are assumed and posterior distributions derived using Gibbs sampling. These are used to identify differentially expressed genes. A similar approach has recently been proposed to relate genes/markers to disease for genome-wide association studies [10]. As an alternative to Gibbs sampling, which is computationally intense, the Bayesian score criterion [11] can be used as an approximation to the posterior probability of the network, and the significance of the pathway may be assessed on the basis of a randomization test of the score [12].

An alternative network-based approach [13] represents a network as a directed acyclic graph to which a latent variable (or random effect) for each gene representing its baseline level is added, allowing a mixed effect model for the expression levels to be derived. This allows inference to be based on linear mixed effect model techniques. For example, tests of treatment effect on a node set can be derived as Wald tests, making use of standard approximation methods.

Methods exploiting pathway structure have also been proposed for other, related purposes. For classification, knowledge of an undirected gene network has been used to develop classifiers of gene profiles by performing a spectral decomposition of the expression profiles with respect to the eigenfunctions of the graph [14]. For prediction, a modified Lasso algorithm using a penalty function that takes pathway structure into account has been proposed [15]. Similarly, a modified boosting algorithm which incorporates the pathway structure has been proposed [16].

In the following section we describe a simple way to incorporate a known network or pathway into the analysis of gene expression data. This entails augmenting the network with a discrete node, representing the treatment or class variable. We show that this leads to a simple modification of conventional differential expression analysis. The augmented network contains a discrete as well as multiple continuous (Gaussian) nodes: networks containing both types of node are usually called hybrid networks. With a few recent exceptions [17-20], hybrid networks have been little used in network biology. They may be useful, for example, when modeling expression data from comparative experiments (with discrete design variables) or when seeking to integrate such data into a wider biological context (with discrete genotypic or phenotypic information). Since hybrid networks may be unfamiliar to some readers we describe the framework carefully in the next section, but stress that no novel theory is involved.

Methods

The model framework

We suppose that data from a gene expression study are available, in the form of an $N \times p$ data matrix $X = \{x_g : g \in V\}$, where N is the number of observations and V is a set of p genes, together with an $N \times 1$ vector T containing the treatment group information. The key assumption we make is that the steady state (or unperturbed) distribution of the gene expression levels follows a given Bayesian network model, specified in the form of a DAG (directed acyclic graph) $\mathcal{G} = (V, E)$, where E is a set of directed edges between the nodes in V. This network is assumed to have been obtained from external sources, as we illustrate below.

Under such a model, the joint distribution of the data $X = \{x_g : g \in V\}$ factorizes into a product of conditional distributions

$$f(X) = \prod_{g \in V} f\left(x_g \mid x_{\mathrm{pa}_{\mathcal{G}}(g)}\right) \tag{1}$$

where $\mathrm{pa}_{\mathcal{G}}(g)$ is the parents of node $g \in V$ in $\mathcal{G}$, that is, the set of nodes with an arrow to g in $\mathcal{G}$. One of the strengths of Bayesian network methodology is the freedom to use any univariate models for the conditional distributions. In this paper we assume Gaussian models, in which the conditional distributions are linear regressions of x_g with covariates given by the variables $x_{\mathrm{pa}_{\mathcal{G}}(g)}$: the Gaussian assumption is not critical for the approach, as we further discuss below.

Another strength of the methodology is the ability to read from the graph which conditional independences hold under the model, using the property of d-separation [21,22]. For disjoint variable sets $U, V, W \subseteq V$ we write

$$x_U \perp\!\!\!\perp x_V | x_W$$

to mean that x_U and x_V are conditionally independent given x_W.

To model the effect of the treatment or class variable T, we add a node labeled T to V, and suppose that the treatment directly affects a set of genes $V_T \subseteq V$. Thus we consider DAGs of the form

$$\mathcal{G}(E) = (V \cup \{T\}, E \cup E_T)$$

where E_T is a set of additional edges of the form (T, g) for $g \in V_T$. We suppose that the object of the analysis is to find (i.e., estimate) V_T.

We assume that $\mathcal{G}(E)$ defines a Bayesian network model for (X, T). Since $\mathcal{G}(E)$ contains both discrete and Gaussian nodes, it is a hybrid Bayesian network [23]. From (1) we have that

$$\begin{aligned} f(X') &= \prod_{y \in V \cup \{T\}} f\left(x_y \mid x_{\mathrm{pa}_{\mathcal{G}(E)}(g)}\right) \\ &= f(x_T) \left[\prod_{g \in V} f(x_g \mid x_{\mathrm{pa}_{\mathcal{G}(E)}(g)}) \right] \end{aligned} \tag{2}$$

Comparing (1) with (2) we see two changes: firstly, a term $f(x_T) = \Pr(x_T)$ is introduced. Since we usually condition on T this term is of little interest. Secondly, for those genes g in V_T we need to let the conditional distribution of x_g depend on x_T as well as $x_{\mathrm{pa}_{\mathcal{G}}(g)}$. The simplest way to do this is to include an additive treatment effect term, corresponding to a simple shift in the conditional distributions, so that the coefficients of the remaining covariates $x_{\mathrm{pa}_{\mathcal{G}}(g)}$ do not depend on the treatment. This implies that the treatment may affect the mean gene expression values but not their covariances. In some applications it may be more appropriate to let the coefficients vary by treatment, by specifying treatment by covariate interactions, but in the following we assume additive treatment effects.

Maximum likelihood estimates under the model (2) can be obtained by maximizing the likelihood for each factor separately: since these are all standard models, this is easily done. The likelihood ratio test (or *deviance*) statistic for testing $\mathcal{G}(E_0)$ under $\mathcal{G}(E_1)$ when $E_0 \subseteq E_1$ is $2(\hat{l}_0 - \hat{l}_1)$, where $\hat{l}_0$ and $\hat{l}_1$ are the maximized log-likelihood values under $\mathcal{G}(E_0)$ and $\mathcal{G}(E_1)$. Under $\mathcal{G}(E_0)$ the deviance has an asymptotic χ^2_f distribution where$f$ is the difference in the number of parameters of the two models.

An important special case occurs when E_0 and E_1 differ by one edge only, say $T \to g$. Observe that in (2), removing (T,g) only affects the component $f(x_g \mid x_{\mathrm{pa}_{\mathcal{G}(E)}(g)})$. Thus a test for the removal of the edge $T \to g$ from $\mathcal{G}(E_1)$ is a test of

$$x_g \perp\!\!\!\perp x_T|x_S \text{ versus } x_g \not\perp\!\!\!\perp x_T|x_S \qquad (3)$$

where $S = \mathrm{pa}_{\mathcal{G}}(g)$. That is, it is a test of the hypothesis, say $\mathcal{H}_g$, that x_g is conditionally independent of x_T given the parents of g in $\mathcal{G}$. It is natural to evaluate such a test in the conditional distribution of x_g given $x_{S\cup\{T\}}$ [22, p.110-2]. This follows a linear regression model for x_g with covariates x_S and a discrete term for x_T, and so to test $\mathcal{H}_g$ we test whether the coefficient(s) of the discrete term are zero in this model. This is a classical F-test (or when T is binary, a t-test) which is valid in small samples under the usual distributional assumptions. Since it does not depend on which other edges are in E_1, we can associate each edge $T \to g$ with an F-test. This is in contrast to, for example, variable selection in linear models where evidence for the presence or absence of a term depends on which other terms are taken to be present in the model.

Testing the conditional independence of each gene and treatment, given the parents of the gene in the network can be regarded as a simple modification of conventional methods for differential expression analysis that are based on tests of marginal independence between treatment and genes. To compare and contrast the conditional and marginal approaches, consider the two models relating a treatment T to the expression levels of two genes, g_1 and g_2 shown in Figure 1. Conditioning increases power when there is a direct treatment effect, and reduces type II error when there is an indirect but no direct effect, so making the inference more precise.

Note also that the marginal approach can be regarded as the special case of the conditional approach that occurs when $\mathcal{G} = (V, \emptyset)$, so that $\mathrm{pa}_{\mathcal{G}}(g) = \emptyset$ for all $g \in V$.

Since multiple hypotheses are tested, use of conventional significance level thresholds would inflate the false positive rate. Many approaches to correct for multiplicity are available [24]. When the number of genes in the network is large, control of the false discovery rate may be desirable. However, when the number of genes is small, most methods for control of false discovery rate perform badly [25], and other methods have been suggested [26,27]. One option is to use a classical multiple test procedure such as Holm's step-down procedure [28], as we illustrate in the next section.

Results

In this section we describe two applications of the method.

Hepatotoxicity and the PPAR pathway

Here we describe the analysis of data taken from a hepatotoxicity study in rats (*Rattus norvegicus*) available from the Gene Expression Omnibus [29], a public data repository under the auspices of the National Center for Biological Information. The study has ID GSE24363.

The stated objective of the study was to use microarray gene expression data acquired from the liver of rats exposed to hepatotoxicants to build classifiers for prediction of liver necrosis. In the study 418 rats were exposed to one of eight compounds (1,2-dichlorobenzene, 1,4-dichlorobenzene, bromobenzene, monocrotaline,

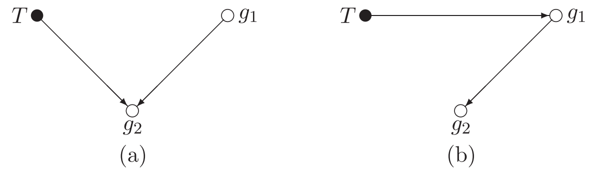

Figure 1 Comparison of marginal and conditional tests. Comparison of conditional and marginal tests for two models. Under (**a**), where T has a direct effect on g_2, $T \not\perp\!\!\!\perp g_2$ and $T \not\perp\!\!\!\perp g_2 \mid g_1$, but the conditional test will generally have greater power than the marginal test, since using g_1 as a regressor will explain some proportion of g_2's variation. Under (**b**), where T does not have a direct effect on g_2, $T \not\perp\!\!\!\perp g_2$ but $T \perp\!\!\!\perp g_2 \mid g_1$. Hence the conditional null hypothesis holds, and the Type II error of the conditional test is less than α, the Type I error of the test. In contrast, the marginal hypothesis is more likely to be rejected, and the Type II error is inflated.

N-nitrosomorpholine, thioacetamide, galactosamine, and diquat dibromide). All eight compounds were studied using standardized procedures, i.e. a common array platform (Affymetrix Rat 230 2.0 microarray), experimental procedures and data retrieving and analysis processes. For each compound, four to six male, 12 week old F344 rats were exposed to a zero dose, low dose, mid dose(s) or a high dose of the toxicant and sacrificed at 6, 24 or 48 hrs later. At necropsy, liver was harvested for RNA extraction, histopathology, and clinical chemistry assessments.

For simplicity we use the subset of data from the study pertaining to 1,2-dichlorobenzene, and compare active with control treatments (ignoring the effects of dose and exposure time). The preprocessing steps are described on the GEO website. In all there were 46 arrays in the subset: 12 animals were in the control group, and 34 animals were exposed to active drug.

Peroxisome proliferator-activated receptors (PPARs) are a group of nuclear receptor proteins that function as transcription factors, playing essential roles in the regulation

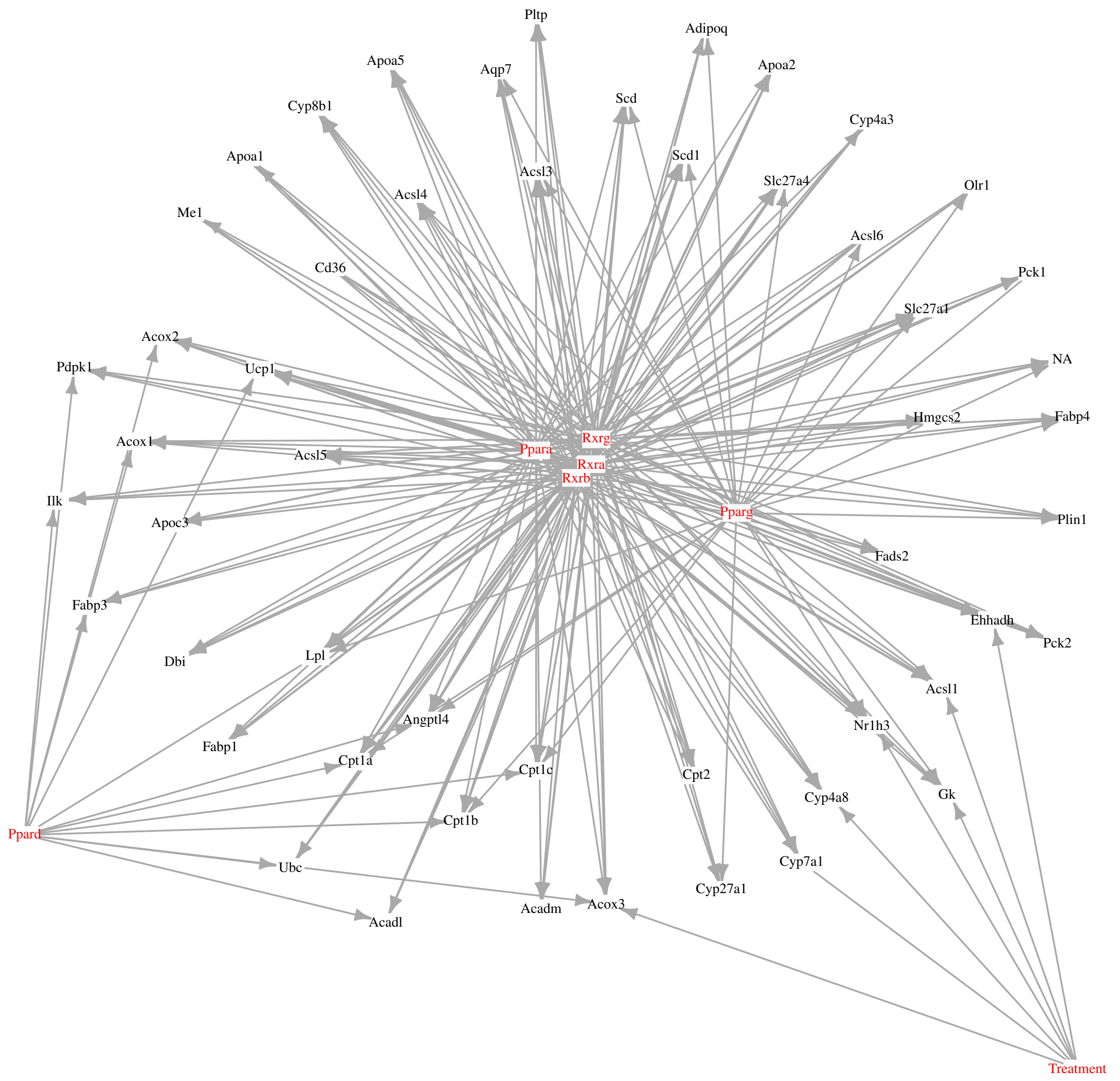

Figure 2 Augmented PPAR pathway. An inferred PPAR pathway showing the effects of treatment (multiplicity-adjusted p-values less than 0.05). Transcription factors are shown in red.

of cellular differentiation, development, and metabolism of higher organisms. Several types have been identified, denoted PPAR-α, PPAR-δ and PPAR-γ. The signalling pathway controlling the expression of the PPARs and their targets has been well-studied. We are interested in studying the effects of the hepatotoxicant on this pathway.

We obtained a copy of the KEGG [5] PPAR signalling pathway for *Rattus norvegicus* using the SPIA package [30]. All edges correspond to transcription factor/ target gene relations. Excluding genes that are not present on the array, we obtain a DAG with 78 nodes and 287 edges, which is used in the analysis.

To examine the effect of treatment on the network, the network-based tests of $x_T \perp\!\!\!\perp x_g|x_{\text{pa}(g)}$ for each gene g were compared to marginal t-tests, that is, of $x_T \perp\!\!\!\perp x_g$. In both cases, to correct for multiplicity we use Holm's step-down procedure [28] to obtain simultaneous (i.e., multiplicity-adjusted) p-values, denoted $\{\tilde{p_g}, g \in V\}$. If we reject precisely those hypotheses $\mathcal{H}_g$ for which $\tilde{p}_g < \alpha$, the family-wise error rate is controlled, i.e. the probability that this procedure results in *any* false rejections is less or equal to α. This is true irrespective of which hypotheses in $\mathcal{H} = \{\mathcal{H}_g, g \in V\}$ are true or false. Here false rejections equate to false edge inclusions. This is a cautious approach, that aims to identify the edges which are certainly present in E: see [31].

Using the network-based tests, $\tilde{p} < 0.05$ for seven genes, but none using the marginal tests. Figure 2 shows the augmented PPAR pathway showing the seven genes affected by treatment. Using the same criteria, the marginal analysis detects no treatment effects. Thus in this example, conditioning on the parents of each gene in the network results in a substantial increase in power.

The effects of smoking

Here we describe the analysis of data taken from a study of the effects of cigarette smoke on the human oral mucosal transcriptome [32]. These are available from the Gene Expression Omnibus [29]: the study has ID GSE17913. In all 40 current smokers and 40 age- and gender- matched never smokers participated in the study, the primary objective of which was to determine the effects of smoking on gene expression in oral epithelial cells. One subject was excluded from the analysis due to poor data quality. For further information on the study and data preprocessing we refer to [32] and [29].

Here we use the data to characterize the effect of smoking on gene expression, making use of a global transcription network constructed using information on human transcription factors (TFs) and their putative target genes (TGs) obtained from the TRANSFAC database [33]. Using the GSEABase package [34] we downloaded data on 95919 (TF,TG) pairs for in all 615 putative TFs and 8945 putative TGs. Of these, the expression levels of 7932 genes were recorded in the GSE17913 study. The (TF,TG) pairs for these genes generate a directed graph, say $\mathcal{G}_0$, with 7932 nodes (of which 191 are TFs), and 61982 edges. Since some TFs are targets of other TFs, $\mathcal{G}_0$ contains cycles. The subgraph $\mathcal{G}_1$ of $\mathcal{G}_0$ induced by the TFs has 191 nodes and 2711 edges. We derived an acyclic subgraph of $\mathcal{G}_1$ by omitting edges with the least strength of evidence, in the following way. To derive a measure for the strength of evidence for each edge in $\mathcal{G}_1$, we fit a linear regression model for the expression of each gene, including its parents in the network as covariates, and use standard significance tests for zero coefficients as a measure of strength of evidence. Starting with the null graph we added edges, ordered in terms of decreasing strength of evidence (increasing p-value), that do not generate cycles. This resulted in an acyclic subgraph of $\mathcal{G}_1$, say $\mathcal{G}_2$, with 191 nodes and 2110 edges. We then added the nodes and edges in $\mathcal{G}_0$ but not in $\mathcal{G}_1$ to $\mathcal{G}_2$, resulting in a DAG, say $\mathcal{G}$, with 7932 nodes and 61381 edges.

We applied the method to $\mathcal{G}$, correcting for multiplicity using the method of [35]. Figure 3 compares the adjusted p-values from the analysis with the corresponding adjusted values from the marginal analysis not exploiting the network topology. Clearly the conditional tests have substantially greater power. The numbers of genes satisfying the false discovery rate at 1%, 2.5% and 5% were 123, 254 and 478 respectively. In comparison, for a marginal analysis the corresponding numbers were 25, 60 and 162.

It is instructive to relate the differentially expressed genes to the network topology. For purposes of illustration we consider the 254 genes that satisfy the false

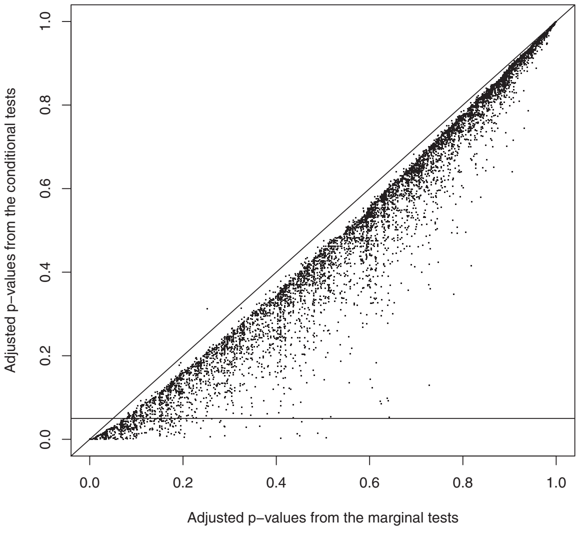

Figure 3 Comparison of marginal and conditional adjusted p-values. A scatterplot of multiplicity-adjusted p-values obtained using the conditional and tests for the global transcription network.

discovery rate at 2.5%, of which 10 are transcription factors. The subnetwork of $\mathcal{G}$ induced by these 254 genes has one connected component with 107 genes, the remainder being isolated nodes. Figure 4 displays the connected subnetwork, which contains all 10 transcription factors. Thus the analysis has identified a highly connected transcriptional module that is strongly affected by smoking.

Discussion

We have described a simple way to exploit network information in the analysis of gene expression data, using tests for the conditional independence of each gene and treatment given the parents of the gene in the network. This method can be regarded as an extension of conventional methods of gene expression analysis, that takes the network structure into account. We demonstrated using two examples that the method can result in a substantial increase in power.

In a related approach to the analysis of genetics of gene expression data [36], when relating a marker to the expression of a gene, all expression levels were considered as potential regressors; model selection methods were used

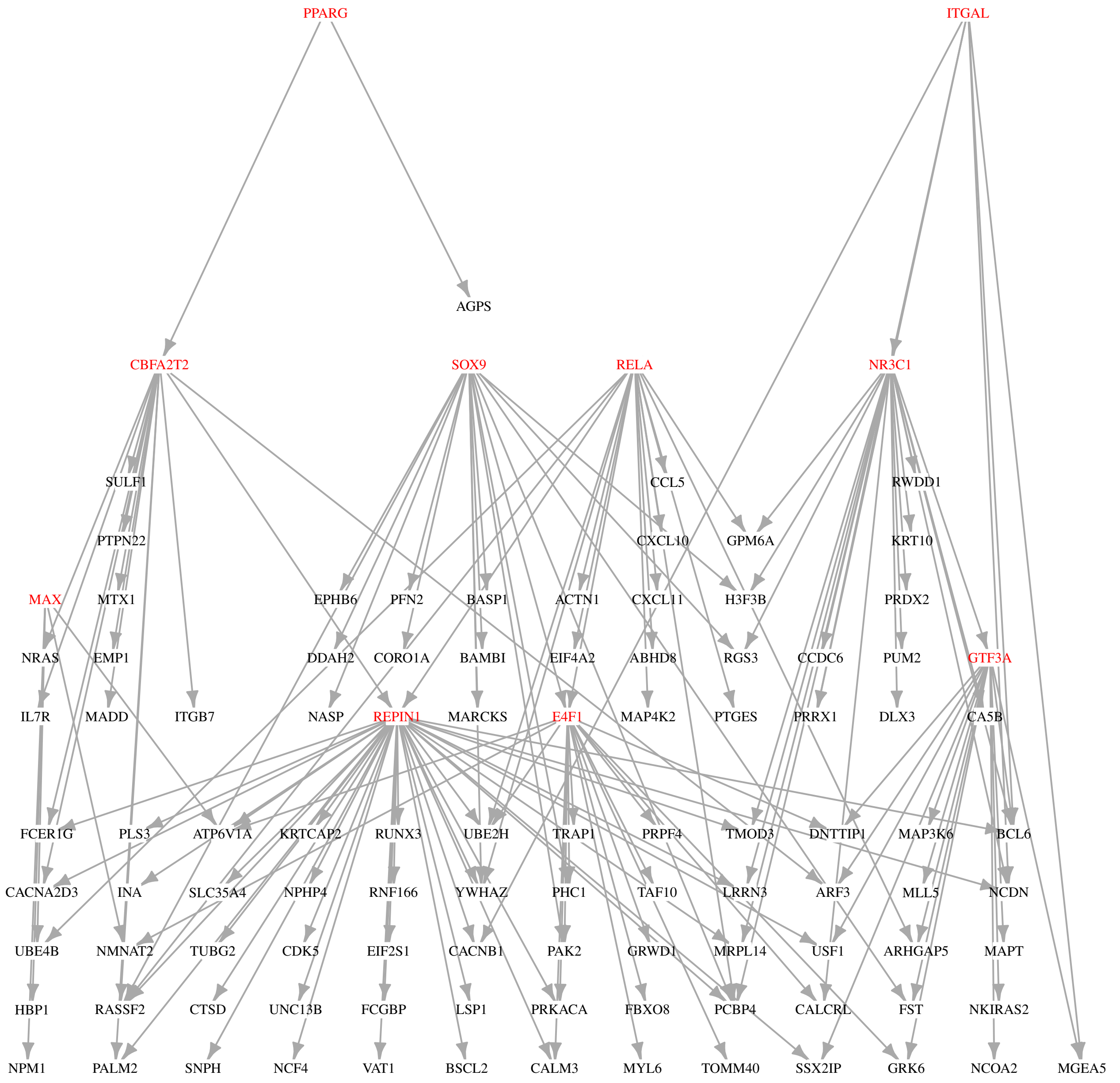

Figure 4 A subnetwork of the global transcription network. A subnetwork of the global transcription network. The expression of all genes in the subnetwork are affected by smoking (false discovery rate of 2.5%). Transcription factors are shown in red.

to choose the regressors for each marker-gene pair. This is in contrast to the method proposed here, in which an external biological network determines the regressors. We remark that if the regressors for a gene under study include some that are causally downstream of the gene, this may dilute or remove evidence of a direct marker effect on the expression of the gene.

As described above, some authors [9,12] incorporate a dichotomous treatment into a network by letting the network nodes represent discretized fold changes rather than expression levels. We prefer our approach for several reasons. Firstly, the use of fold changes does not appear to be biologically motivated. For example, as we discuss below, it is expected that the mRNA abundance of a transcription factor affects the mRNA abundance of its target genes, but there seems no reason for this to be true for the corresponding fold changes. Secondly, our framework allows direct and indirect treatment effects to be distinguished, and thirdly, it may easily be extended to handle multiple treatment variables, although we do not consider such extensions here.

The approach builds on some assumptions that may or may not be unwarranted. The key assumption is that the steady state distribution of the gene expression levels follows a given Bayesian network. Gene regulation is extremely complex and as yet imperfectly understood, so such an assumption can at best be tentative. We have illustrated the approach using two networks, one based on a signalling pathway and the other constructed using transcription factor/target gene data. Biochemical pathways represent phenomena occurring at the protein level, which are not necessarily reflected at the transcript level: see Figure 5. Thus caution is required when using such pathways as models for gene expression data.

As we have described the method, it assumes that the expression data are Gaussian distributed, but this is not critical. Expression data from microarrays are typically taken to be Gaussian after log transformation [37,38]. Other technologies such as RNA-seq generate discrete count data that are better modeled using Poisson or negative binomial distributions [39]. This can be handled in the current framework by using the appropriate univariate conditional models in (2). What is essential in our context is the availability of an appropriate test for (3), that is, for the conditional independence of treatment and the level of expression of a gene, given the expression levels of its parents in the network. If the correct form of the conditional distribution is in doubt, a permutation test may be used.

An assumption, implicit in both the marginal and the network-based analyses, is that the treatment variable is causal rather than reactive in respect to the gene expression data. Ideally the treatment should represent a randomized intervention, allowing secure causal interpretations, but gene expression studies are rarely randomized. In poorly designed studies, treatment allocation may be confounded with other factors [40]. When this is true, the interpretation of edges $T \rightarrow g$ as representing effects of treatment is compromised. The hepatotoxicity study described above is an example of a well-designed study in which treatment represents an non-randomized intervention that is orthogonal to other sources of variation. It is reasonable to interpret treatment effects as causal in such studies. In observational studies, such as the study of the effects of smoking described above, allocation to treatment group (i.e., the process leading to an individual becoming a smoker or non-smoker) could in principle be determined by factors that also influence gene expression, so jeopardizing the interpretation of edges $T \rightarrow g$ as treatment effects. In the actual study, however, it is natural to

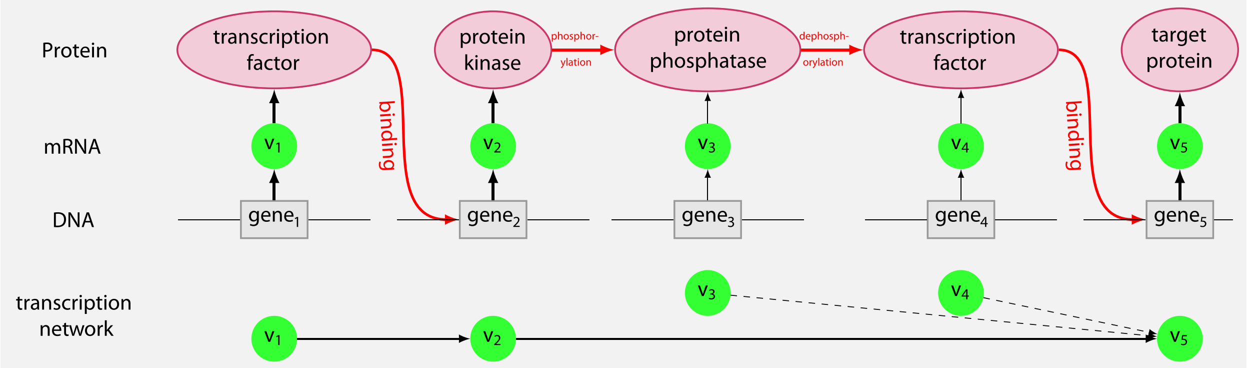

Figure 5 A hypothetical signalling pathway. The figure shows a hypothetical signaling pathway adapted from [41], containing two transcription factors, a protein kinase and a protein phosphatase. Three levels are shown: DNA, mRNA and proteins, as well the expected transcription network. At the transcript level, gene 1 affects gene 2 which affects gene 5, but gene 2 does not affect genes 3 or 4. If proteins 3 and/or 4 are rate-limiting, this will be reflected in the dependences shown as dotted arrows.

interpret edges in terms of the effects of tobacco smoke on epithelial cells.

Similar remarks apply to the use of the terms *direct effects* and *indirect effects*: a causal interpretation of these may be unwarranted.

Finally, it is assumed that the treatment affects the parameters of the network but not its topology. In some applications this may not be appropriate. For example, interventions affecting chromatin structure may alter the accessibility of DNA binding sites and hence patterns of regulatory control due to transcription factors.

Conclusions

A straightforward way to exploit network information in the analysis of gene expression data is to assume that the network models the steady state distribution of the gene expression levels, and that the treatment affects the parameters but not the topology of the network. In this framework, genes whose expression is directly affected by the treatment may be identified using tests for the conditional independence of each gene and treatment given the parents of the gene in the network. This method can be regarded as an extension of conventional methods of gene expression analysis that takes network structure into account. It is simple to implement, gives rise to substantial power gains, and may give insight into the biological processes involved.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The financial support of Quantomics, a collaborative project under the 7th Framework Programme (FP7) is gratefully acknowledged (PS).

Author's contributions

DE conceived the methods, performed the analyses and drafted the manuscript. LW participated in the analyses. All authors read and approved the final manuscript.

Received: 16 November 2011 Accepted: 28 June 2012
Published: 16 July 2012

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**(43):15545–15550. [http://dx.doi.org/10.1073/pnas.0506580102].
2. Ackermann M, Strimmer K: **A general modular framework for gene set enrichment analysis.** *BMC Bioinf* 2009, **10:**47. [http://www.biomedcentral.com/1471-2105/10/47].
3. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis A, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25–29.
4. Emmert-Streib F, Glazko GV: **Pathway analysis of expression data: deciphering functional building blocks of complex diseases.** *PLoS Comput Biol* 2011, **7**(5):e1002053. [http://dx.doi.org/10.1371%2Fjournal.pcbi.1002053].
5. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27–30.
6. Rahnenfuhrer J, Domingues F, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data.** *Stat Appl in Genet and Mol Biol* 2004, **3:**Article 16.
7. Draghici S, Khatri P, Tarca A, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis.** *Genome Res* 2007, **17:**1537–1545.
8. Geistlinger L, Csaba G, Küffner R, Mulder N, Zimmer R: **From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems.** *Bioinf* 2011, **27:**i366–i373.
9. Wei Z, Li H: **A Markov random field model for network-based analysis of genomic data.** *Bioinformatics* 2007, **23**(12):1537–1544. [http://bioinformatics.oxfordjournals.org/content/23/12/1537.abstract].
10. Chen M, Cho J, Zhao H: **Incorporating biological pathways via a markov random field model in genome-wide association studies.** *PLoS Genet* 2011, **7**(4):e1001353. [http://dx.doi.org/10.1371%2Fjournal.pgen.1001353].
11. Heckerman D, Geiger D, Chickering DM: **Learning bayesian networks: The combination of knowledge and statistical data.** *Machine Learning* 1995, **20:**197–243.
12. Isci S, Ozturk C, Jones J, Otu HH: **Pathway analysis of high-throughput biological data within a Bayesian network framework.** *Bioinf* 2011, **27**(12):1667–1674. [http://bioinformatics.oxfordjournals.org/content/27/12/1667.abstract].
13. Shojaie A, Michailidis G: **Analysis of gene sets based on the underlying regulatory network.** *J Comput Biol* 2009, **16:**407–426.
14. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP: **Classification of microarray data using gene networks.** *BMC Bioinf* 2007, **8:**35. [http://www.biomedcentral.com/1471-2105/8/35].
15. Li C, Li H: **Network-constrained regularization and variable selection for analysis of genomic data.** *Bioinf* 2008, **24**(9):1175–1182.
16. Binder H, Schumacher M: **Incorporating pathway information into boosting estimation of high-dimensional risk prediction models.** *BMC Bioinf* 2009, **10:**18. [http://www.biomedcentral.com/1471-2105/10/18].
17. Neto EC, Keller MP, Attie AD, Yandell BS: **Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetics architecture for correlated phenotypes.** *Ann Appl Stat* 2010, **4:**320–339.
18. Kiiveri H: **Multivariate analysis of microarray data: differential expression and differential connection.** *BMC Bioinf* 2011, **12:**42. [http://www.biomedcentral.com/1471-2105/12/42].
19. Pérez A, Larrañaga P, Inza I: **Supervised classification with conditional gaussian networks: increasing the structure complexity from naive bayes.** *Int J Approximate Reasoning* 2006, **43:**1–25. [http://www.sciencedirect.com/science/article/pii/S0888613X0600003X].
20. Edwards D, de Abreu GCG, Labouriau R: **Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests.** *BMC Bioinf* 2010, **11:**18.
21. Pearl J: *Probabilistic Inference in Intelligent Systems*. San Mateo, CA: Morgan Kauffmann; 1988.
22. Lauritzen SL: *Graphical Models*. Oxford, UK: Oxford University Press; 1996.
23. Lerner UN: **Hybrid bayesian networks for reasoning about complex systems.** *Tech. rep*:Stanford University, 2002.
24. Dudoit S, van der Laan, M: *Multiple Testing Procedures with Applications to Genomics*. New York: Springer-Verlag; 2008.
25. Efron B: *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge, UK: Cambridge University Press; 2010.
26. Bickel DR: **Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions.** *ArXiv* 2011, **1106:**4490.
27. Bickel DR: **Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison.** *ArXiv* 2011, **1104:**0341.
28. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Stat* 1979, **6:**65–70.
29. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207–210.
30. Tarca AL, Khatri P, Draghici S: **Bioconductor's SPIA package.** 2010. [bioconductor.org/packages/2.6/bioc/vignettes/SPIA/inst/doc/SPIA.pdf].
31. Drton M, Perlman MD: **A SINful approach to Gaussian graphical model selection.** *J Stat Plan Inf* 2008, **138:**1179–1200.

32. Boyle JO, Gümüş ZH, Kacker A, Choksi VL, Bocker JM, Zhou XK, Yantiss RK, Hughes DB, Du B, Judson BL, Subbaramaiah K, Dannenberg AJ: **Effects of cigarette smoke on the human oral mucosal transcriptome.** *Cancer Prev Res* 2010, **3:**266–278.
33. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28:**316–319. [http://nar.oxfordjournals.org/content/28/1/316.abstract].
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**(43):15545–15550. [http://dx.doi.org/10.1073/pnas.0506580102].
35. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J R Stat Soc Ser B (Methodological)* 1995, **57:**289–300. [http://www.jstor.org/stable/2346101].
36. Perez-Enciso M, Quevedo J, Bahamonde A: **Genetical Genomics: use all data.** *BMC genomics* 2007, **8**(69):8–69.
37. Allison DB, Cui X, Page GP: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7:**1471–0056.
38. Smyth G: **Linear models and empirical Bayes for assessing differential expression in microarry experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**.
39. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results.** *Genome Biol* 2010, **11:**220.
40. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nat Rev Genet* 2010, **11:**733–739.
41. Wagner A: **How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps.** *Bioinformatics* 2001, **17**(12):1183–1197.

doi:10.1186/1471-2105-13-167
Cite this article as: Edwards *et al.*: Network-enabled gene expression analysis. *BMC Bioinformatics* 2012 **13**:167.